

# ***KanScan***

***Version 1.0b***

***User's Guide***

## Contents

<b>1</b>	<b>WHAT IS KANSCAN?.....</b>	<b>3</b>
1.1	INTRODUCING KANSCAN .....	3
<b>2</b>	<b>INSTALLING KANSCAN.....</b>	<b>5</b>
2.1	KANSCAN CLIENT WINDOW .....	5
<b>3</b>	<b>KANSCAN USER INTERFACE.....</b>	<b>7</b>
3.1	SELECTING THE FILE TO BE DIGITIZED - FILE MENU .....	7
3.2	INCREASING THE SIZE OF THE IMAGE – SCALE IMAGE .....	10
3.3	HORIZONTAL AND VERTICAL SCROLL BARS.....	11
3.4	CONVERTING THE GREY SCALE IMAGE INTO A BINARY IMAGE - BINARIZE.....	11
3.5	FINDING LETTER'S BODY HEIGHT.....	14
3.6	FINDING DISTANCE BETWEEN TEXT BLOCKS .....	15
3.7	FINDING THE MINIMUM HORIZONTAL DISTANCE BETWEEN 2 ADJACENT TEXT BLOCKS.....	15
3.8	FINDING THE MINIMUM VERTICAL DISTANCE BETWEEN 2 ADJACENT TEXT BLOCKS.....	16
3.9	DIGITIZING THE DOCUMENT IMAGE .....	17
<b>4</b>	<b>HELP, SUPPORT AND FEEDBACK .....</b>	<b>19</b>

## 1 What is KanScan?

### 1.1 Introducing KanScan

There is abundant information and literature in Kannada language on paper documents. This format does not allow for an easy process of sharing, searching, editing or re-creation. Each of these tasks will need huge investments in manual labor and time. Most of such tasks are normally abandoned or not taken up at all, due to the prohibitive costs.

The task of converting the information in paper documents into editable documents is called *digitization*. An OCR (Optical Character Recognition) solution is the need of the hour, which can automate the digitization process. While there are many tools commercially available in the market for English and European languages, there is none available for Kannada language so far. **KanScan** software is an attempt to bridge that gap.

**KanScan** is an intelligent OCR software for recognizing text in images of printed documents or books, containing Kannada language text and creating editable text documents,

Digitizing your Kannada printed documents using KanScan will enable you to harness the information which you were so far not able to access.

An online version of the software is available for free to instantly convert individual image pages into editable files.

If you have a pile of documents or book(s) to be digitized, you can contact us. KanScan can easily process book pages or many documents in batch.

#### KanScan Features

- OCR software for digitizing Kannada printed documents or books
- Font independent recognition  
The OCR algorithms used are not based on fonts. Instead, these algorithms look for distinguishing character features to recognize characters. This means KanScan is not limited to one or few particular fonts.
- Fast processing  
Each document is processed in a few seconds
- Transparent and Automated pre-processing  
KanScan detects and applies corrections or adjustments to input images like deskewing.
- Accurate recognition

A good quality document image can result in recognition accuracy of more than 90%.

- Recognition of pictures within documents  
If there is a picture embedded between text blocks, KanScan can identify the picture from the surrounding text and ignore the picture elements.
- Ease of use  
KanScan's very simple and user friendly interface, which allows you to use the software quickly, easily and effectively.

There are two components in the KanScan software. The *KanScan Client* and *KanScan OCR Server*.

You need to download KanScan client application, to start using KanScan. The client application allows user to

- select an image from the user's computer
- utilize the user interface to provide some basic input that is required by KanScan
- send the image to the remote KanScan OCR Server
- receive the digitized document from the remote KanScan OCR Server

KanScan OCR server resides on a remote machine, on a website. It is necessary that your network connection is available, so that the client can send document images to the KanScan OCR Server.

The following sections in the document will provide you instructions on how to install the KanScan client and use it to digitize Kannada printed documents.

## 2 Installing KanScan

It is very likely that you have already installed KanScan, based on the instructions that were on the website. You may skip this section.

The following are required before you install KanScan:

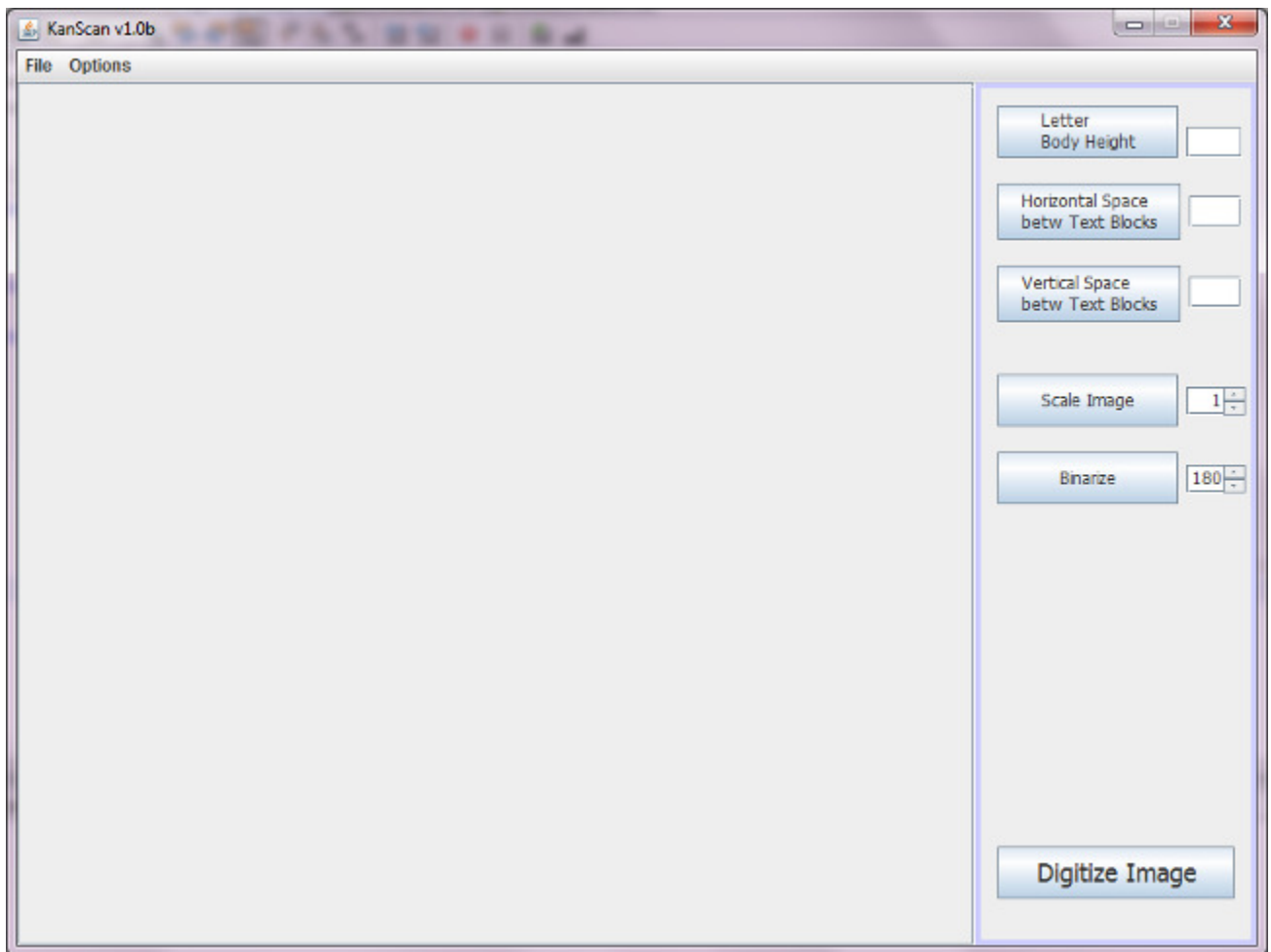
- PC with Microsoft XP, Win 7, or MS Vista operating system
- Java Runtime Environment

In case, you have not installed KanScan client, then:

- Make a folder named *KanScan* on your computer
- Go to site <http://www.kannadaocr.com/>. This is the Home page of KanScan.
- Click on the link – [Download](#). You will now be in the *KanScan Download* page.
- On the left side column of the page, Click on the link [KanScan Client download](#).
- Save the file in the KanScan folder you had created.

### 2.1 KanScan Client Window

To bring up the KanScan application, double-click on the file *kanscan10b.jar* in the KanScan folder. This will open up the KanScan user interface client window. A screen snapshot is given below:



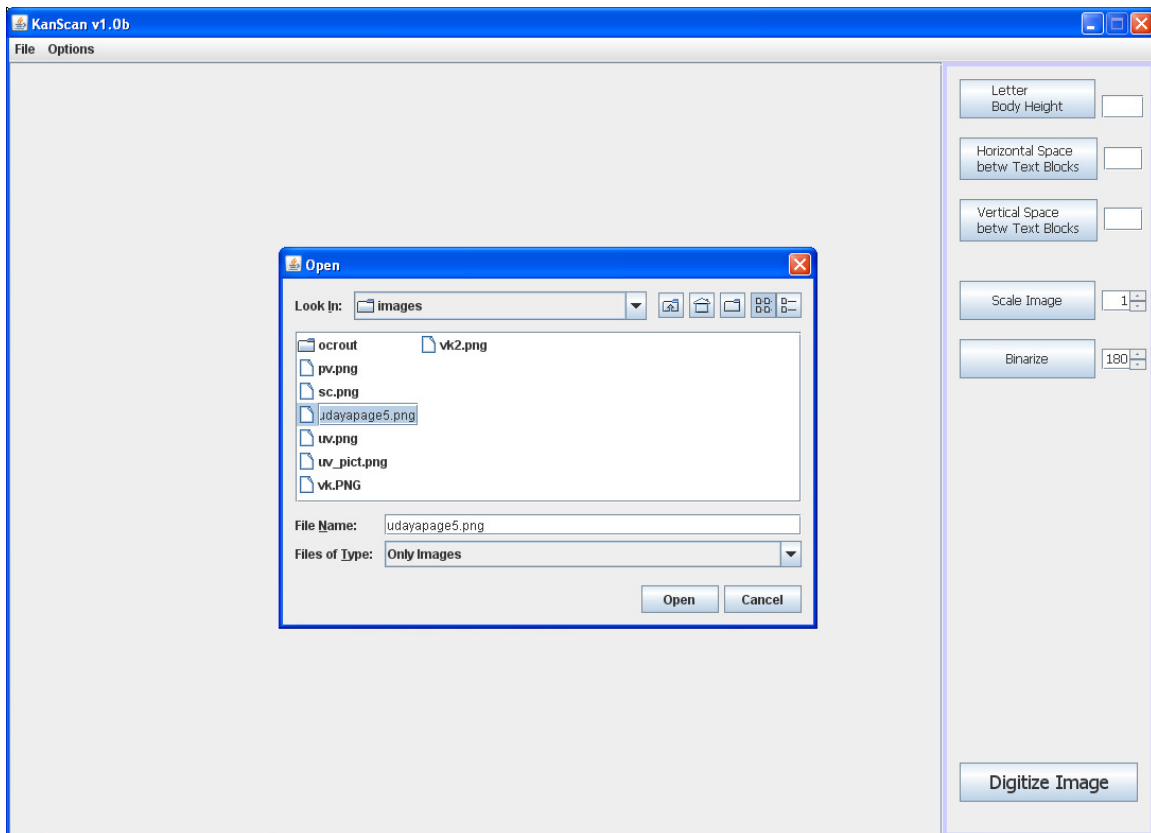
### 3 KanScan User Interface

KanScan has a very simple, user friendly interface, which enables user to start using the software very quickly.

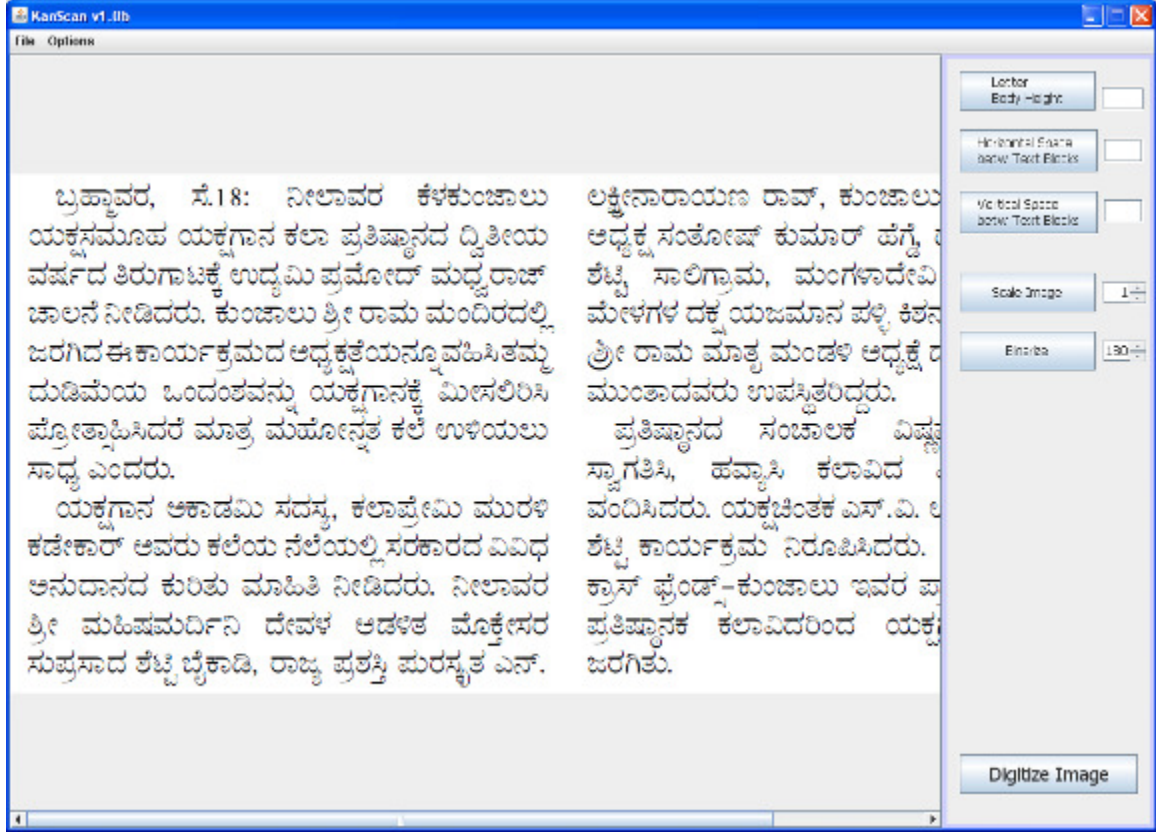
#### 3.1 Selecting the file to be digitized - File menu

Let us assume that you have an image named 'udayapage5.png', that you want to digitize.

- Click on the File menu on the menu bar.
- This will show a File **Open** pop-up window.



- Go to the folder where the image is stored and choose the image file udayapage5 .png.
- Click on **Open** button.
- The image is loaded and shown on the KanScan window.



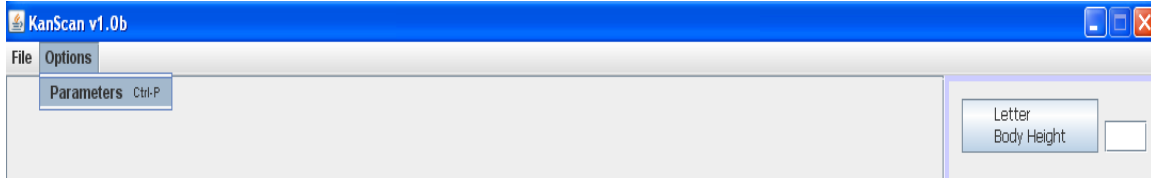


### *Setting the default folder for image files*

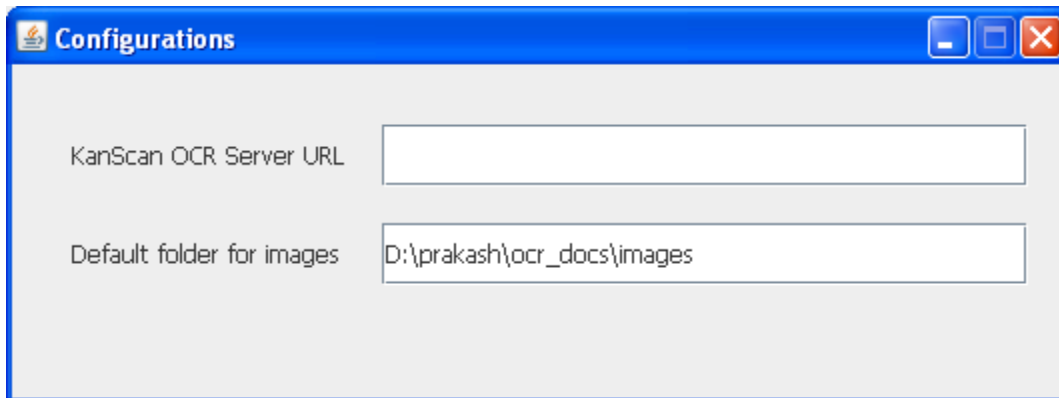
Your document images may all be located in one particular folder. In case, you want the *File Open* pop-up window to always show you this folder contents, you can setup this folder as default folder.

Here are the steps to do that:

1. Click on menu **Options**, and the submenu item **Parameters**.



The **Configurations** screen is displayed.



2. In the text box beside **Default folder for images**, enter the default folder path for your images.
3. Close the pop-up box by clicking the X.

### 3.2 Increasing the size of the image – Scale Image

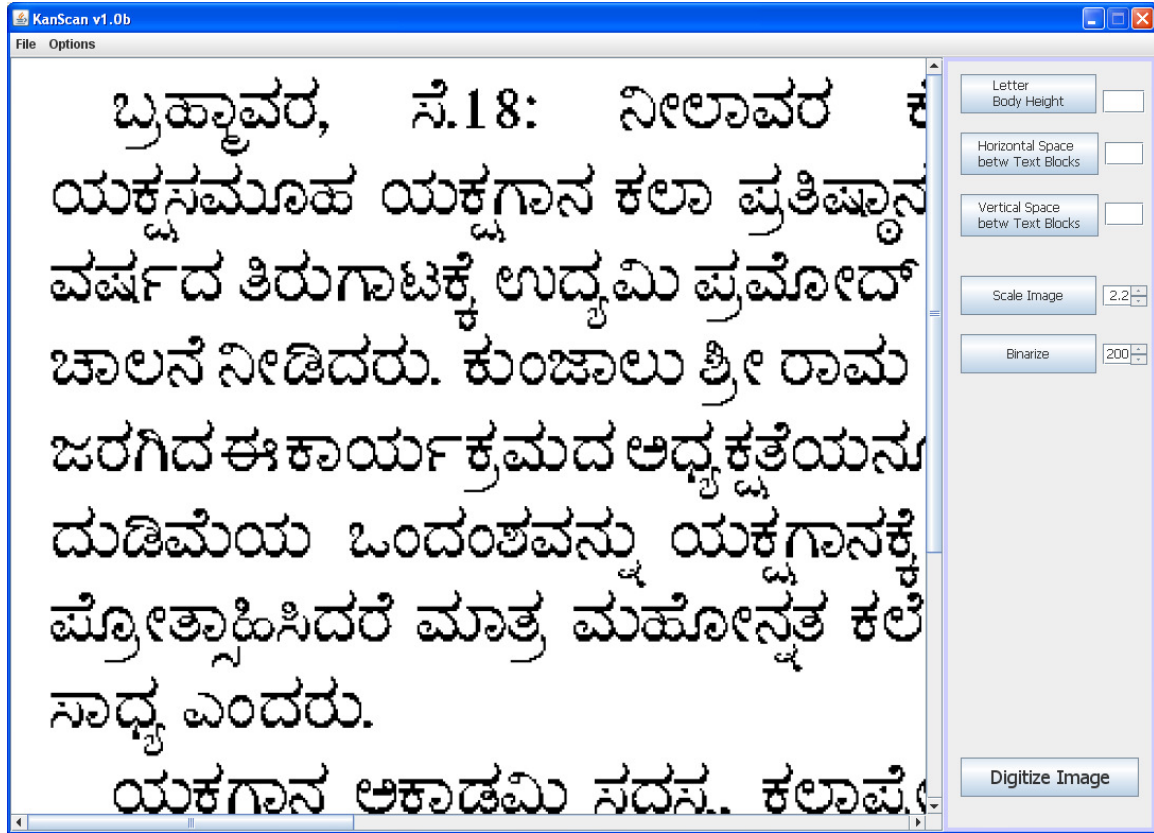
If size of letters is small, it is better to increase the size of the image. This will help in accurately measuring the letter height, horizontal and vertical space of the text.

You can increase or decrease the size of the image by specifying a *scale factor* and clicking on the **Scale Image** button.

The text letters in the above example are legible but little small for accurately measuring the height of the text line. Let us increase the image size by a scale factor of 2.2, so as to increase the size of the image by 2.2 (increase by 220%) in comparison to the original loaded image.

- Enter 2.2 in the text box beside the button **Scale Image**.
- Click on the Scale image button.

The scaled image appears as below:



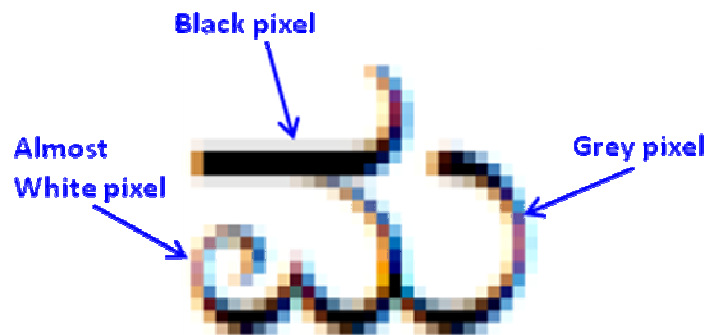
### 3.3 Horizontal and Vertical Scroll bars

Note that because image size is bigger than the KanScan window, part of the image is not visible.

Use the vertical and horizontal scroll bars to move the image vertically and horizontally and view the hidden parts.

### 3.4 Converting the Grey Scale image into a Binary Image - Binarize

The scanned image of your text is not actually a black & white image, but, rather has different shades of grey, ranging from black to white. If you increase the image and observe each letter closely, you will notice the different grey pixels that make up the letter.



Digitization begins with first converting the grey-scale image into a binary image. It is much easier to work with a binary image.

The conversion is done by choosing a '*binarization threshold*' value, and then

- converting all pixels with value less than this threshold value to a white pixels
- converting all pixels with value greater or equal to this threshold value to black pixels

The binary image will now have only black & white pixels.

You need to provide the threshold value to KanScan, by choosing that value in the input box beside the button named **Binarize**. You will notice that the values range from 50 to 240.

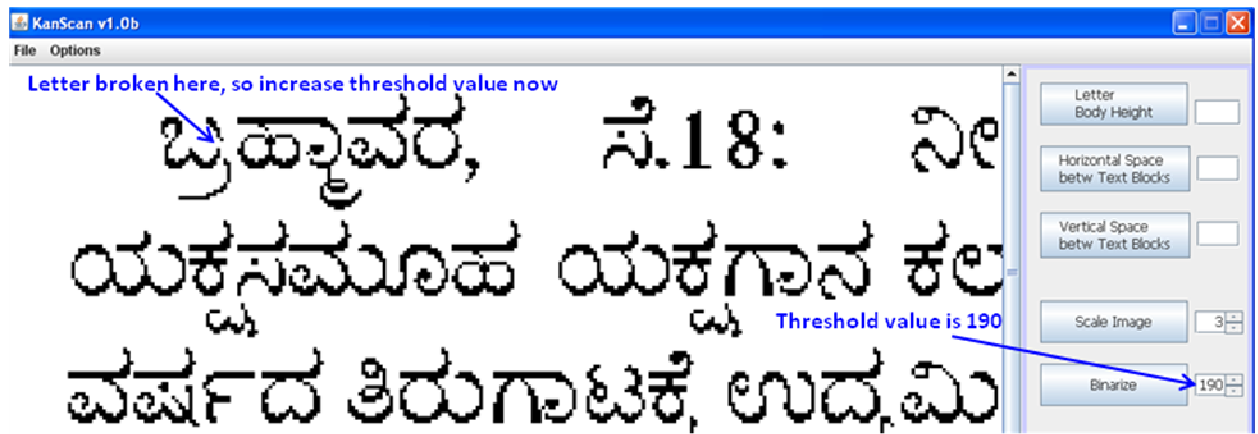
How do you choose the correct Binarization value? Here are the steps:

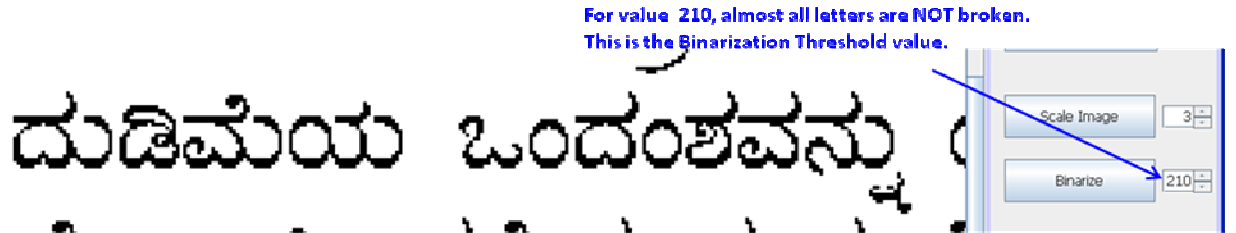
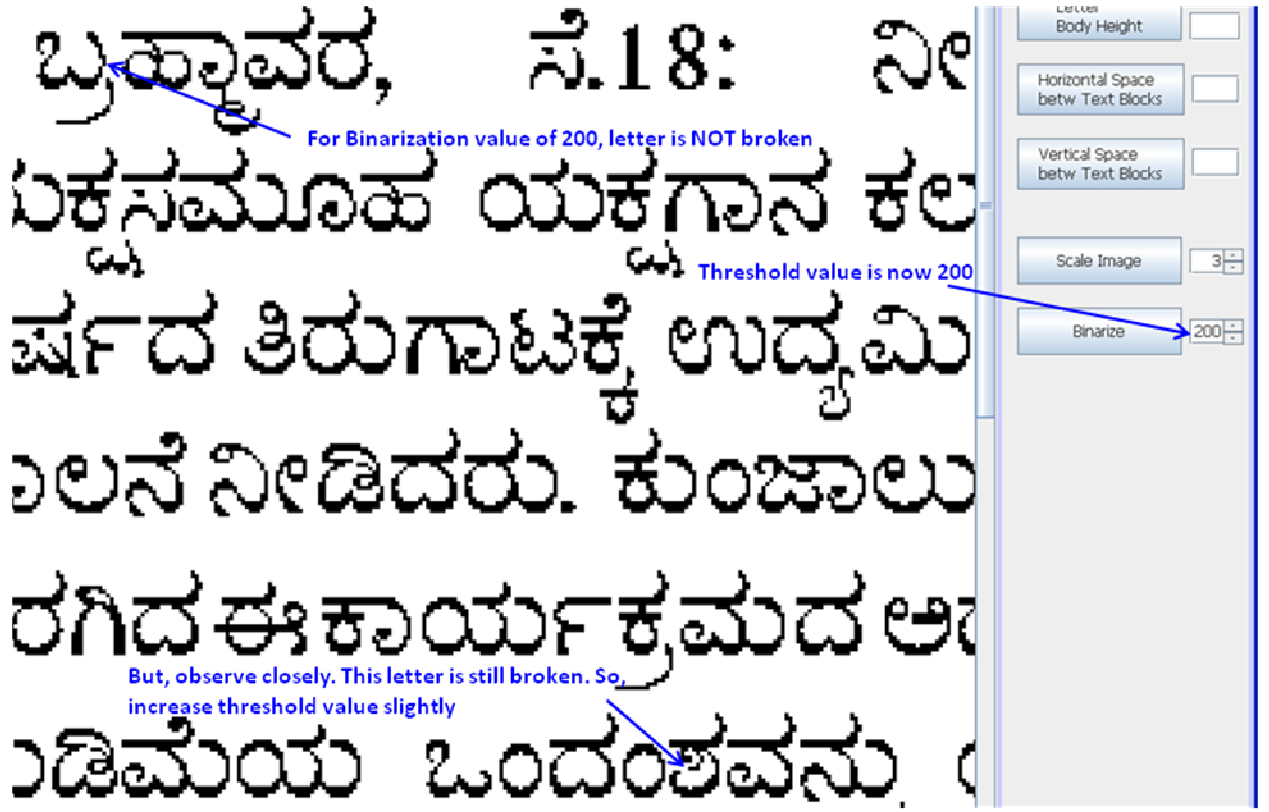
1. If the letters in the window are small, increase the view size by using the Scale Image feature, described in earlier section.
2. Choose a value of 190 for binarization threshold and click on the **Binarize** button.

KanScan will now show the binarized image in the window. You will notice that the grey pixels have disappeared or have been converted into black pixels.

3. Observe each letter closely.  
Binarization will change some grey-pixels into white pixel. This may cause the letter to be broken. The pixels making up a letter, which were earlier all connected, may now appear disconnected, with the absence of black or grey pixel(s).
4. If none of the letters are broken, then, decrease the binarization value. Else, if some letters are broken, then increase the binarization value. Click on **Binarize** button.
5. Repeat steps 3 & 4 until you find a threshold value, which does not cause the letter to be broken, but, any slight increase in this value will result in broken letters. This would be the *Binarization Threshold Value*.

Figure below shows the text image for different Binarization values.





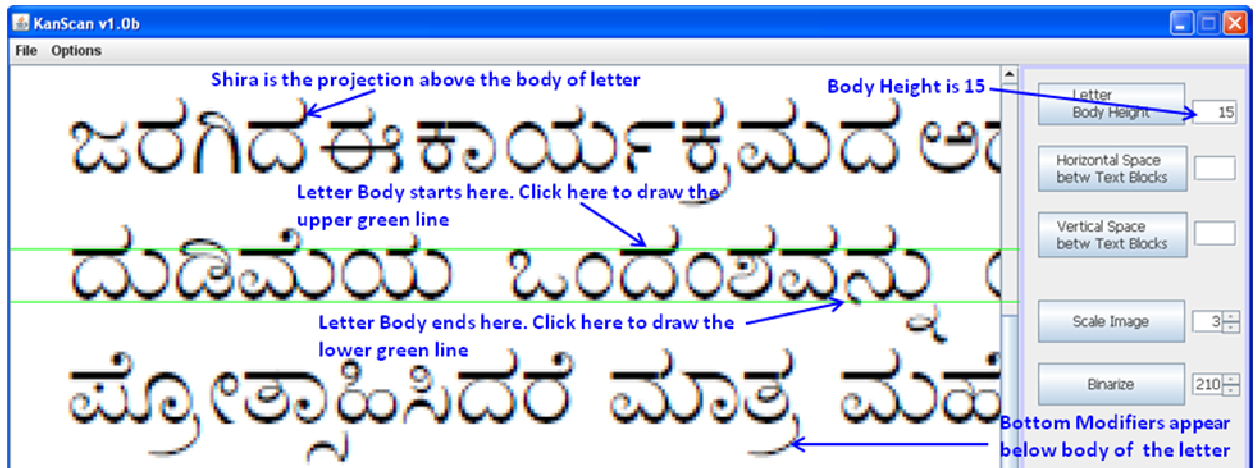
### 3.5 Finding Letter's Body Height

We now have to find the letter body height. A document may contain text blocks which have different height. For example, the 'header' text and 'matter' text could be of different height. But, what we are interested in is the text that makes up the bulk of the document, the 'matter' text.

The body of a letter is the middle part of the letter. It will be below the *Shira* of the letter. The *Shira* is the projection above the body of the letter.

The *Bottom Modifiers* of a letter are the accents that appear below the body of the letter.

1. Make sure that the letters are sufficiently large. Increase the size of the image by entering a scale factor and clicking the **Scale Image** button.
2. Hold together the **SHIFT** key and **UP ARROW** key.
3. Move the mouse cursor to the body start of the letter.
4. Click on the LEFT mouse button. A green line is now drawn, indicating the start of the letter body.
5. If you find that the line is a little away from the body start, repeat step 2 to 4, until you are satisfied that the green line is exactly at the start of body. You have now identified the start of body of the letter.
6. Hold together the **SHIFT** key and **DOWN ARROW** key.
7. Move the mouse cursor to the body end of the letter.
8. Click on the LEFT mouse button. A green line is now drawn, indicating the end of the letter body.
9. If you find that the line is a little away from the body end, repeat step 6 to 8, until you are satisfied that the green line is exactly at the end of body. You have now identified the end of body of the letter.
10. The body height is the number of pixels between the two green lines you have drawn just now. Click on the **Letter Body Height** button. You should now see the height of the body in the text box beside the button. In the example, the height was calculated to be 15.



### 3.6 Finding Distance between Text Blocks

Content in a document is organized into blocks of text and images. These blocks are usually rectangular and sometimes free-form.

The figure below shows a document which has rectangular blocks.

It is important for OCR software to recognize text blocks and process them individually. Otherwise, the output from OCR would be a series of text lines, without any logical sequence.

KanScan will be able to recognize blocks that are rectangular. It is also able to distinguish between text and image blocks.

*Horizontal Space between Text Blocks* is the distance in pixels between two horizontally adjacent text blocks.

*Vertical Space between Text Blocks* is the distance in pixels between two vertically adjacent text blocks.

The horizontal (or vertical) distance, sometimes, may not be the same, for all two horizontally (or vertically) adjacent text blocks.

KanScan is interested in the shortest or minimum, horizontal (or vertical) distance between text blocks for the document.

KanScan is able to automatically detect text blocks and separate them, if the horizontal (and vertical) distance is above 25 pixels, even if you do not specify these inputs.

In case, the distance is lesser than 25 pixels, you should let KanScan know the minimum distances.

The following 2 sections will tell you how to find these values.

### 3.7 Finding the Minimum Horizontal distance between 2 adjacent Text Blocks

Here are the steps:

#### *Drawing the Left Vertical line*

1. Look at the document image and locate the 2 text blocks whose horizontal distance is lesser than any other 2 adjacent text blocks
2. Make sure that the letters are sufficiently large. Increase the size of the image by entering a scale factor and clicking the **Scale Image** button.
3. Scroll, if required to bring the located text blocks into the window.
4. Hold down the **LEFT ARROW** key.
5. Move the mouse cursor very close to the end of the left text block.
6. When you are sure that the cursor is just about the end of the text block, click on the **LEFT MOUSE** button. Note that all the while the LEFT ARROW key is pressed down.
7. A vertical line is shown at the point where you clicked. You can now release the LEFT ARROW key.
8. The left text box should completely be on the left side of the vertical line. If it is not, then, repeat steps 3 to 7

#### *Drawing the Right Vertical line*

1. Hold down the **RIGHT ARROW** key.
2. Move the mouse cursor very close to the beginning of the right text block.
3. When you are sure that the cursor is just about the beginning of the text block, click on the **LEFT MOUSE** button. Note that all the while the RIGHT ARROW key is pressed down.
4. A vertical line is shown at the point where you clicked. You can now release the RIGHT ARROW key.
5. The left text box should completely be on the left side of the vertical line. If it is not, then, repeat steps 1 to 4

#### *Horizontal Distance between Text Blocks*

1. Now click on the **Horizontal Space betw Text Blocks** button. The horizontal distance in pixels is shown in text-box beside the button.

---

### **3.8 Finding the Minimum Vertical distance between 2 adjacent Text Blocks**

Here are the steps:

#### *Drawing the Top Horizontal line*

1. Look at the document image and locate the 2 text blocks whose vertical distance is lesser than any other 2 adjacent text blocks
2. Make sure that the letters are sufficiently large. Increase the size of the image by entering a scale factor and clicking the **Scale Image** button.
3. Scroll, if required to bring the located text blocks into the window.
4. Hold down the **UP ARROW** key.
5. Move the mouse cursor very close to the end of the upper text block.
6. When you are sure that the cursor is just about the end of the text block, click on the **LEFT MOUSE** button. Note that all the while the UP ARROW key is pressed down.
7. A horizontal line is shown at the point where you clicked. You can now release the UP ARROW key.
8. The upper text box should completely be on the top of the horizontal line. If it is not, then, repeat steps 3 to 7

#### *Drawing the Bottom Vertical line*

1. Hold down the **DOWN ARROW** key.
2. Move the mouse cursor very close to the top of the lower text block.
3. When you are sure that the cursor is just about the top of the text block, click on the **LEFT MOUSE** button. Note that all the while the DOWN ARROW key is pressed down.
4. A horizontal line is shown at the point where you clicked. You can now release the DOWN ARROW key.
5. The lower text box should completely be on the lower side of the horizontal line. If it is not, then, repeat steps 1 to 4



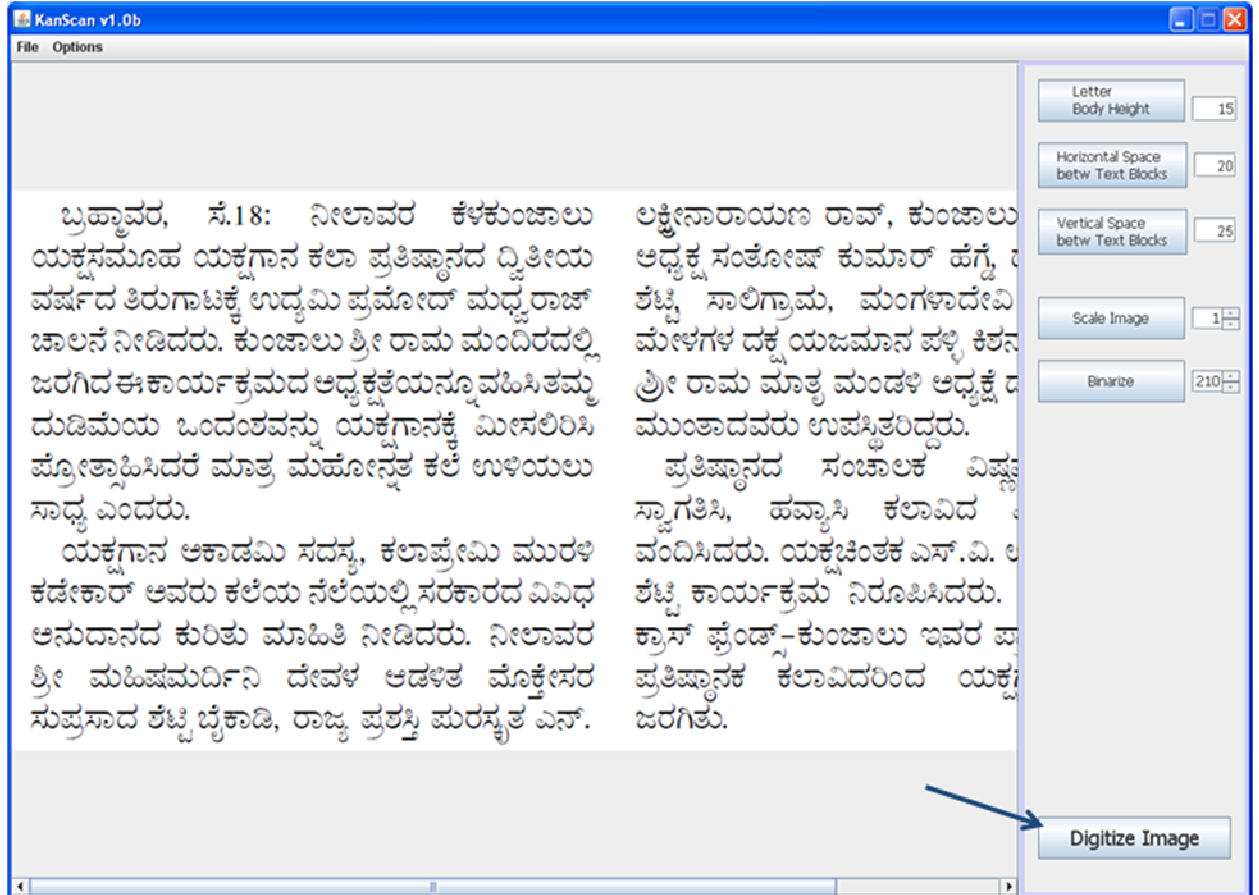
### Vertical Distance between Text Blocks

1. Now click on the **Vertical Space betw Text Blocks** button. The vertical distance in pixels is shown in text-box beside the button.

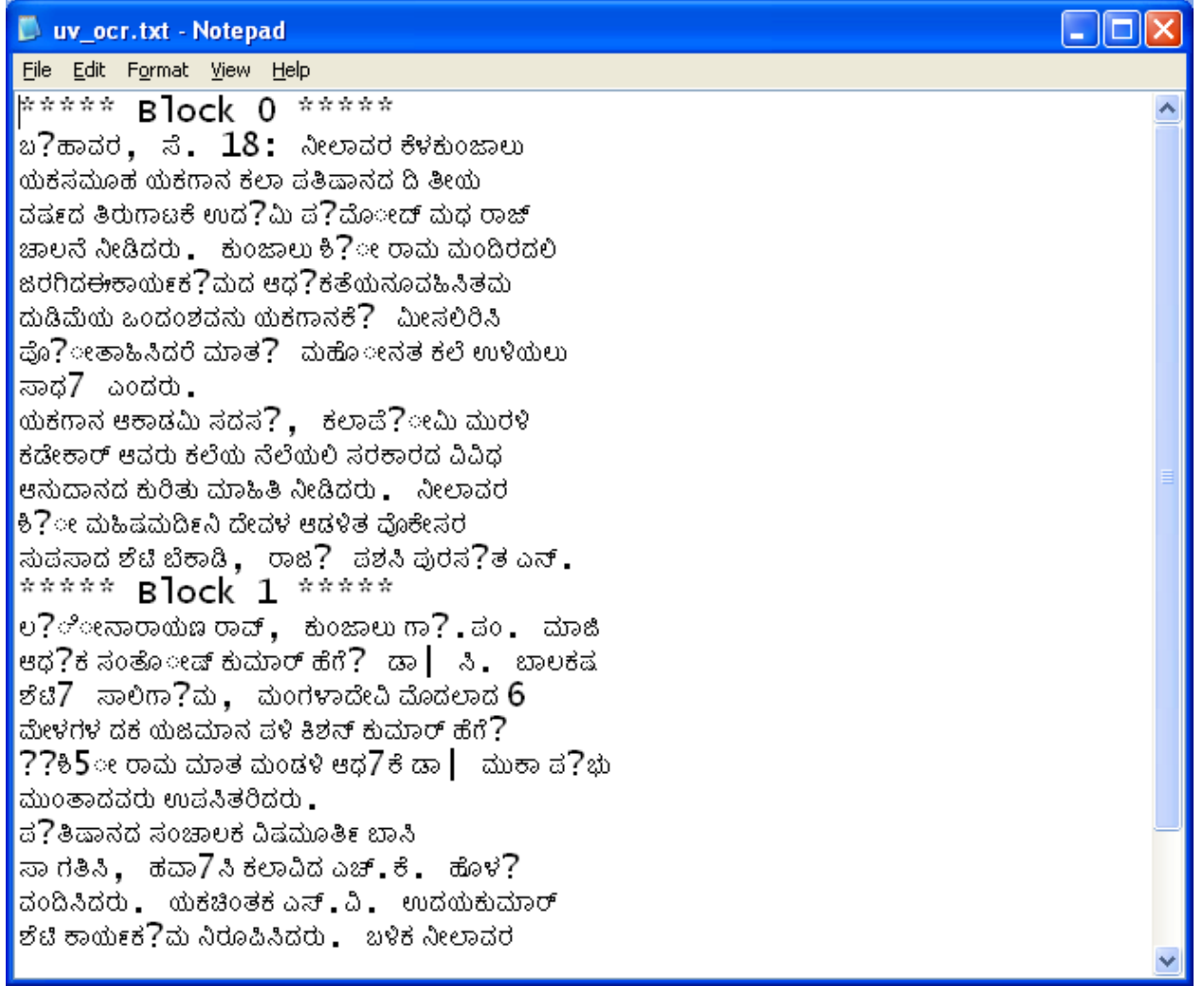
## 3.9 Digitizing the Document Image

Now that all the above steps are completed, you are ready to digitize the document image. You need to send the document image to the KanScan OCR server to get the digitized file.

1. Make sure that internet connection is active.
2. Click on the **Digitize Image** button.



3. The image document is sent to KanScan OCR server. A few seconds later the digitized output is shown as a text file in Notepad application.
4. You can save this text file on your computer.



---

## **4 Help, Support and Feedback**

In case you have any issues or you need help or want to send feedback, you can send an email [bvprakash@gmail.com](mailto:bvprakash@gmail.com)

\*\*\* End of Document \*\*\*